

Transductive Learning with Multi-class Volume Approximation

Gang Niu

Tokyo Institute of Technology, Japan
gang@sg.cs.titech.ac.jp

Bo Dai

Georgia Institute of Technology, USA
bohr.dai@gmail.com

Marthinus Christoffel du Plessis

Tokyo Institute of Technology, Japan
christo@sg.cs.titech.ac.jp

Masashi Sugiyama

Tokyo Institute of Technology, Japan
sugi@cs.titech.ac.jp

Abstract

Given a hypothesis space, the *large volume principle* by Vladimir Vapnik prioritizes equivalence classes according to their volume in the hypothesis space. The *volume approximation* has hitherto been successfully applied to binary learning problems. In this paper, we extend it naturally to a more general definition which can be applied to several transductive problem settings, such as *multi-class*, *multi-label* and *serendipitous* learning. Even though the resultant learning method involves a non-convex optimization problem, the globally optimal solution is almost surely unique and can be obtained in $O(n^3)$ time. We theoretically provide stability and error analyses for the proposed method, and then experimentally show that it is promising.

1 Introduction

The history of the *large volume principle* (LVP) goes back to the early age of the statistical learning theory when Vapnik (1982) introduced it for the case of hyperplanes. But it did not gain much attention until a creative approximation was proposed in El-Yaniv et al. (2008) to implement LVP for the case of soft response vectors. From then on, it has been applied to various binary learning problems successfully, such as binary transductive

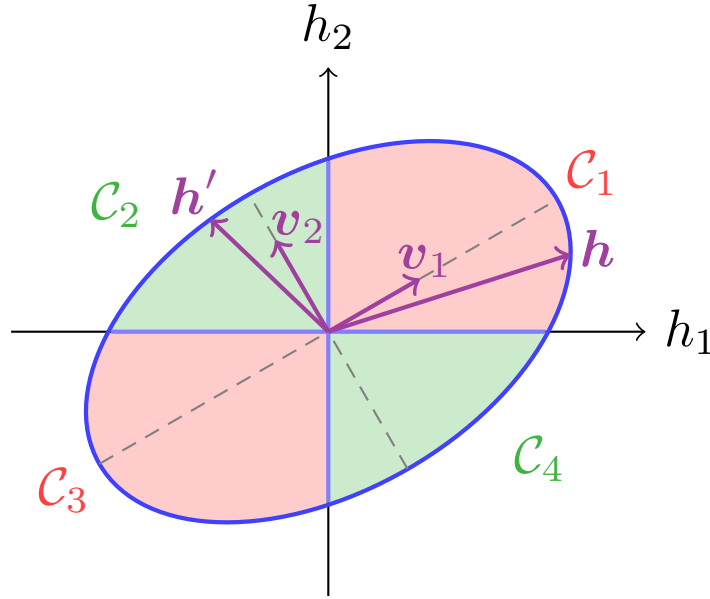


Figure 1: The large volume principle and its approximation.

learning (El-Yaniv et al., 2008), binary clustering (Niu et al., 2013a), and outlier detection (Li and Ng, 2013).

LVP is a learning-theoretic principle which views learning as *hypothesis selecting* from a certain *hypothesis space* \mathcal{H} . Despite the form of the hypothesis, \mathcal{H} can always be partitioned into a finite number of equivalence classes after we observe certain data, where an *equivalence class* is a set of hypotheses that generate the same labeling of the observed data. LVP, as one of the learning-theoretic principles from the statistical learning theory, prioritizes those equivalence classes according to the volume they occupy in \mathcal{H} . See the illustration in Figure 1: The blue ellipse represents \mathcal{H} , and it is partitioned into $\mathcal{C}_1, \dots, \mathcal{C}_4$ each occupying a quadrant of the Cartesian coordinate system \mathbb{R}^2 intersected with \mathcal{H} ; LVP claims that \mathcal{C}_1 and \mathcal{C}_3 are more preferable than \mathcal{C}_2 and \mathcal{C}_4 , since \mathcal{C}_1 and \mathcal{C}_3 have larger volume than \mathcal{C}_2 and \mathcal{C}_4 .

In practice, the hypothesis space \mathcal{H} cannot be as simple as in Figure 1. It frequently locates in very high-dimensional spaces where exact or even quantifiable *volume estimation* is challenging. Therefore, El-Yaniv et al. (2008) proposed a *volume approximation* to bypass the volume estimation. Instead of focusing on the equivalence classes of \mathcal{H} , it directly focuses on the hypotheses in \mathcal{H} since learning is regarded as hypothesis selecting in LVP. It defines \mathcal{H} via an ellipsoid, measures the angles from hypotheses to the principal axes of \mathcal{H} , and then prefers hypotheses near the long principal axes to those near the short ones. This manner is reasonable, since the long principal axes of \mathcal{H} lie in large-volume regions. In Figure 1, \mathbf{h} and \mathbf{h}' are two hypotheses and $\mathbf{v}_1/\mathbf{v}_2$ is the long/short principal axis; LVP advocates that \mathbf{h} is more preferable than \mathbf{h}' as \mathbf{h} is close to \mathbf{v}_1 and \mathbf{h}' is close to \mathbf{v}_2 . We can adopt this volume approximation to regularize our loss function, which has been demonstrated helpful for various binary learning problems.

Nevertheless, the volume approximation in El-Yaniv et al. (2008) only fits binary learning problem settings in spite of its potential advantages. In this paper, we naturally extend it to a more general definition that can be applied to some transductive problem settings including but not limited to *multi-class* learning (Zhou et al., 2003), *multi-label* learning (Kong et al., 2013), and *serendipitous* learning (Zhang et al., 2011). We adopt the same strategy as El-Yaniv et al. (2008): For n data and c labels, a hypothesis space is defined in $\mathbb{R}^{n \times c}$ and linked to an ellipsoid in \mathbb{R}^{nc} , such that the equivalence classes and the volume approximation can be defined accordingly. Similarly to the binary volume approximation, our approach is also distribution free, that is, the labeled and unlabeled data do not necessarily share the same marginal distribution. This advantage of transductive learning over (semi-supervised) inductive learning is especially useful for serendipitous problems where the labeled and unlabeled data must not be identically distributed.

We name the learning method which realizes the proposed multi-class volume approximation *multi-class approximate volume regularization* (MAVR). It involves a non-convex optimization problem, but the globally optimal solution is almost surely unique and accessible in $O(n^3)$ time following Forsythe and Golub (1965). Moreover, we theoretically provide stability and error analyses for MAVR, as well as experimentally compare it to two state-of-the-art methods in Zhou et al. (2003) and Belkin et al. (2006) using USPS, MNIST, 20Newsgroups and Isolet.

The rest of this paper is organized as follows. In Section 2 the binary volume approximation is reviewed, and in Section 3 the multi-class volume approximation is derived. In Section 4, we develop and analyze MAVR. Finally, the experimental results are in Section 5.

2 Binary Volume Approximation

The binary volume approximation in El-Yaniv et al. (2008) involves a few key concepts: The soft response vector, the hypothesis space and the equivalence class, and the power and volume of equivalence classes. We review the concepts in this section for later use in the next section.

Suppose that \mathcal{X} is the domain of input data, and most often but not necessarily, $\mathcal{X} \subset \mathbb{R}^d$ where d is a natural number. Given a set of n data $X_n = \{x_1, \dots, x_n\}$ where $x_i \in \mathcal{X}$, a *soft response vector* is an n -dimensional vector

$$\mathbf{h} := (h_1, \dots, h_n)^\top \in \mathbb{R}^n, \quad (1)$$

so that h_i stands for a soft or confidence-rated label of x_i . For binary transductive learning problems, a soft response vector \mathbf{h} suggests that x_i is from the positive class if $h_i > 0$, x_i is from the negative class if $h_i < 0$, and the above two cases are equally possible if $h_i = 0$.

A *hypothesis space* is a collection of hypotheses. The volume approximation requires a symmetric positive-definite matrix $Q \in \mathbb{R}^{n \times n}$ which contains the pairwise information about X_n . Consider the hypothesis space

$$\mathcal{H}_Q := \{\mathbf{h} \mid \mathbf{h}^\top Q \mathbf{h} \leq 1\}, \quad (2)$$

where the hypotheses are soft response vectors. The set of sign vectors $\{\text{sign}(\mathbf{h}) \mid \mathbf{h} \in \mathcal{H}_Q\}$ contains all of $N = 2^n$ possible dichotomies of X_n , and \mathcal{H}_Q can be partitioned into a finite number of *equivalence classes* $\mathcal{C}_1, \dots, \mathcal{C}_N$, such that for fixed k , all hypotheses in \mathcal{C}_k will generate the same labeling of X_n .

Then, in statistical learning theory, the *power* of an equivalence class \mathcal{C}_k is defined as the probability mass of all hypotheses in it (Vapnik, 1998, p. 708), i.e.,

$$\mathcal{P}(\mathcal{C}_k) := \int_{\mathcal{C}_k} p(\mathbf{h}) d\mathbf{h}, \quad k = 1, \dots, N,$$

where $p(\mathbf{h})$ is the underlying probability density of \mathbf{h} over \mathcal{H}_Q . The hypotheses in \mathcal{C}_k which has a large power should be preferred according to Vapnik (1998).

When no specific domain knowledge is available (i.e., $p(\mathbf{h})$ is unknown), it would be natural to assume the continuous uniform distribution $p(\mathbf{h}) = 1/\sum_{k=1}^N \mathcal{V}(\mathcal{C}_k)$, where

$$\mathcal{V}(\mathcal{C}_k) := \int_{\mathcal{C}_k} d\mathbf{h}, \quad k = 1, \dots, N,$$

is the *volume* of \mathcal{C}_k . That is, the volume of an equivalence class is defined as the geometric volume of all hypotheses in it. As a result, $\mathcal{P}(\mathcal{C}_k)$ is proportional to $\mathcal{V}(\mathcal{C}_k)$, and the larger the value $\mathcal{V}(\mathcal{C}_k)$ is, the more confident we are of the hypotheses chosen from \mathcal{C}_k .

However, it is very hard to accurately compute the geometric volume of even a single convex body in \mathbb{R}^n , let alone all 2^n convex bodies, so El-Yaniv et al. (2008) introduced an efficient approximation. Let $\lambda_1 \leq \dots \leq \lambda_n$ be the eigenvalues of Q , and $\mathbf{v}_1, \dots, \mathbf{v}_n$ be the associated orthonormal eigenvectors. Actually, the hypothesis space \mathcal{H}_Q in Eq. (2) is geometrically an origin-centered ellipsoid in \mathbb{R}^n with \mathbf{v}_i and $1/\sqrt{\lambda_i}$ as the direction and length of its i -th principal axis. Note that a small angle from a hypothesis \mathbf{h} in \mathcal{C}_k to some \mathbf{v}_i with a small/large index i (i.e., a long/short principal axis) implies that $\mathcal{V}(\mathcal{C}_k)$ is large/small (cf. Figure 1). Based on this crucial observation, we define

$$V(\mathbf{h}) := \sum_{i=1}^n \lambda_i \left(\frac{\mathbf{h}^\top \mathbf{v}_i}{\|\mathbf{h}\|_2} \right)^2 = \frac{\mathbf{h}^\top Q \mathbf{h}}{\|\mathbf{h}\|_2^2}, \quad (3)$$

where $\mathbf{h}^\top \mathbf{v}_i / \|\mathbf{h}\|_2$ means the cosine of the angle between \mathbf{h} and \mathbf{v}_i . We subsequently expect $V(\mathbf{h})$ to be small when \mathbf{h} lies in a large-volume equivalence class, and conversely to be large when \mathbf{h} lies in a small-volume equivalence class.

3 Multi-class Volume Approximation

In this section, we propose a more general multi-class volume approximation that fits for several problem settings.

3.1 Problem settings

Recall the setting of binary transductive problems (Vapnik, 1998, p. 341). A fixed set $X_n = \{x_1, \dots, x_n\}$ of n points from \mathcal{X} is observed, and the labels $y_1, \dots, y_n \in \{-1, +1\}$ of these points are also fixed but unknown. A subset $X_l \subset X_n$ of size l is picked uniformly at random, and then y_i is revealed if $x_i \in X_l$. We call $S_l = \{(x_i, y_i) \mid x_i \in X_l\}$ the labeled data and $X_u = X_n \setminus X_l$ the unlabeled data. Using S_l and X_u , the goal is to predict y_i of $x_i \in X_u$ (while any unobserved $x \in \mathcal{X} \setminus X_n$ is currently left out of account).

A slight modification suffices to extend the setting. Instead of $y_1, \dots, y_n \in \{-1, +1\}$, we assume that $y_1, \dots, y_n \in \mathcal{Y}$ where $\mathcal{Y} = \{1, \dots, c\}$ is the domain of labels and c is a natural number. Though the binary setting is popular, this multi-class setting has been studied in just a few previous works such as Szummer and Jaakkola (2001) and Zhou et al. (2003). Without loss of generality, we assume that each of the c labels possesses some labeled data.

In addition, it would be a multi-label setting, if $y_1, \dots, y_n \subseteq \mathcal{Y}$ with $\mathcal{Y} = \{1, \dots, c\}$ where each y_i is a label set, or if $y_1, \dots, y_n \in \mathcal{Y}$ with $\mathcal{Y} = \{-1, 0, 1\}^c$ where each y_i is a label vector. To the best of our knowledge, the former setting has been studied only in Kong et al. (2013) and the latter setting has not been studied yet. The latter setting is more general, since the former one requires labeled data to be fully labeled, while the latter one allows labeled data to be partially labeled. A huge challenge of multi-label problems is that some label sets or label vectors might have no labeled data (Kong et al., 2013), since there are 2^c possible label sets and 3^c possible label vectors.

A more challenging serendipitous setting which is a multi-class setting but some labels have no labeled data has been studied in Zhang et al. (2011). Let $\mathcal{Y}_l = \{y_i \mid x_i \in X_l\}$ and $\mathcal{Y}_u = \{y_i \mid x_i \in X_u, y_i \notin \mathcal{Y}_l\}$, then we have $\#\mathcal{Y}_u \geq 1$ where $\#$ measures the cardinality. It is still solvable when $\#\mathcal{Y}_u = 1$ if a special label of outliers is allowed and when $\#\mathcal{Y}_u > 1$ as a combination of classification and clustering problems. Zhang et al. (2011) is the unique previous work which successfully dealt with $\#\mathcal{Y}_u = 2$ and $\#\mathcal{Y}_u = 3$.

3.2 Definitions

The multi-class volume approximation to be proposed can handle all the problem settings discussed so far in a unified manner. In order to extend the binary definitions, we need only to extend the hypothesis and the hypothesis space.

To begin with, we allocate a soft response vector in Eq. (1) for each of the c labels:

$$\mathbf{h}_1 = (h_{1,1}, \dots, h_{n,1})^\top, \dots, \mathbf{h}_c = (h_{1,c}, \dots, h_{n,c})^\top.$$

The value $h_{i,j}$ is a soft or confidence-rated label of x_i concerning the j -th label and it suggests that

- x_i should possess the j -th label, if $h_{i,j} > 0$;
- x_i should not possess the j -th label, if $h_{i,j} < 0$;
- the above two cases are equally possible, if $h_{i,j} = 0$.

For multi-class and serendipitous problems, y_i is predicted by $\hat{y}_i = \arg \max_j h_{i,j}$. For multi-label problems, we need a threshold T_h that is either preset or learned since usually positive and negative labels are imbalanced, and y_i can be predicted by $\hat{y}_i = \{j \mid h_{i,j} \geq T_h\}$; or we can use the label set prediction methods proposed in Kong et al. (2013).

Then, a *soft response matrix* as our transductive hypothesis is an n -by- c matrix defined by

$$H = (\mathbf{h}_1, \dots, \mathbf{h}_c) \in \mathbb{R}^{n \times c}, \quad (4)$$

and a *stacked soft response vector* as an equivalent hypothesis is an nc -dimensional vector defined by

$$\mathbf{h} = \text{vec}(H) = (\mathbf{h}_1^\top, \dots, \mathbf{h}_c^\top)^\top \in \mathbb{R}^{nc},$$

where $\text{vec}(H)$ is the vectorization of H formed by stacking its columns into a single vector.

As the binary definition of the hypothesis space, a symmetric positive-definite matrix $Q \in \mathbb{R}^{n \times n}$ which contains the pairwise information about X_n is provided, and we assume further that a symmetric positive-definite matrix $P \in \mathbb{R}^{c \times c}$ which contains the pairwise information about \mathcal{Y} is available. Consider the hypothesis space

$$\mathcal{H}_{P,Q} := \{H \mid \text{tr}(H^\top Q H P) \leq 1\}, \quad (5)$$

where the hypotheses are soft response matrices. Let $P \otimes Q \in \mathbb{R}^{nc \times nc}$ be the *Kronecker product* of P and Q . Due to the symmetry and the positive definiteness of P and Q , the Kronecker product $P \otimes Q$ is also symmetric and positive definite, and $\mathcal{H}_{P,Q}$ in (5) could be defined equivalently as

$$\mathcal{H}_{P,Q} := \{H \mid \text{vec}(H)^\top (P \otimes Q) \text{vec}(H) \leq 1\}. \quad (6)$$

The equivalence of Eqs. (5) and (6) comes from the fact that $\text{tr}(H^\top Q H P) = \text{vec}(H)^\top (P \otimes Q) \text{vec}(H)$ following the well-known identity (see, e.g., Theorem 13.26 of Laub, 2005)

$$(P^\top \otimes Q) \text{vec}(H) = \text{vec}(Q H P).$$

As a consequence, there is a bijection between $\mathcal{H}_{P,Q}$ and

$$\mathcal{E}_{P,Q} := \{\mathbf{h} \mid \mathbf{h}^\top (P \otimes Q) \mathbf{h} \leq 1\}$$

which is geometrically an origin-centered ellipsoid in \mathbb{R}^{nc} . The set of sign vectors $\{\text{sign}(\mathbf{h}) \mid \mathbf{h} \in \mathcal{E}_{P,Q}\}$ spreads over all the $N = 2^{nc}$ quadrants of \mathbb{R}^{nc} , and thus the set of sign matrices $\{\text{sign}(H) \mid H \in \mathcal{H}_{P,Q}\}$ contains all of N possible dichotomies of $X_n \times \{1, \dots, c\}$. In other words, $\mathcal{H}_{P,Q}$ can be partitioned into N equivalence classes $\mathcal{C}_1, \dots, \mathcal{C}_N$, such that for fixed k , all soft response matrices in \mathcal{C}_k will generate the same labeling of $X_n \times \{1, \dots, c\}$.

The definition of the power is same as before, and so is the definition of the volume:

$$\mathcal{V}(\mathcal{C}_k) := \int_{\mathcal{C}_k} dH, \quad k = 1, \dots, N.$$

Because of the bijection between $\mathcal{H}_{P,Q}$ and $\mathcal{E}_{P,Q}$, $\mathcal{V}(\mathcal{C}_k)$ is likewise the geometric volume of all stacked soft response vectors in the intersection of the k -th quadrant of \mathbb{R}^{nc} and $\mathcal{E}_{P,Q}$. By a similar argument to the definition of $V(\mathbf{h})$, we define

$$V(H) := \frac{\mathbf{h}^\top (P \otimes Q) \mathbf{h}}{\|\mathbf{h}\|_2^2} = \frac{\text{tr}(H^\top Q H P)}{\|H\|_{\text{Fro}}^2}, \quad (7)$$

where $\mathbf{h} = \text{vec}(H)$ and $\|H\|_{\text{Fro}}$ means the Frobenius norm of H . We subsequently expect $V(H)$ to be small when H lies in a large-volume equivalence class, and conversely to be large when H lies in a small-volume equivalence class.

Note that $V(H)$ and $V(\mathbf{h})$ are consistent for binary learning problem settings. We can constrain $\mathbf{h}_1 + \mathbf{h}_2 = \mathbf{0}_n$ if $c = 2$ where $\mathbf{0}_n$ is the all-zero vector in \mathbb{R}^n . Let $P = I_2$ where I_2 is the identity matrix of size 2, then

$$V(H) = \frac{\mathbf{h}_1^\top Q \mathbf{h}_1 + \mathbf{h}_2^\top Q \mathbf{h}_2}{\|\mathbf{h}_1\|_2^2 + \|\mathbf{h}_2\|_2^2} = \frac{\mathbf{h}_1^\top Q \mathbf{h}_1}{\|\mathbf{h}_1\|_2^2} = V(\mathbf{h}_1),$$

which coincides with $V(\mathbf{h})$ defined in Eq. (3). Similarly to $V(\mathbf{h})$, for two soft response matrices H and H' from the same equivalence class, $V(H)$ and $V(H')$ may not necessarily be the same value. In addition, the domain of $V(H)$ could be extended to $\mathbb{R}^{n \times c}$ though the definition of $V(H)$ is originally null for H outside $\mathcal{H}_{P,Q}$.

4 Multi-class Approximate Volume Regularization

The proposed volume approximation motivates a family of new transductive methods taking it as a regularization. We develop and analyze an instantiation in this section whose optimization problem is non-convex but can be solved exactly and efficiently.

4.1 Model

First of all, we define the label indicator matrix $Y \in \mathbb{R}^{n \times c}$ for convenience whose entries can be from either $\{0, 1\}$ or $\{-1, 0, 1\}$ depending on the problem settings and whether negative labels ever appear. Specifically, we can set $Y_{i,j} = 1$ if x_i is labeled to have the j -th label and $Y_{i,j} = 0$ otherwise, or alternatively we can set $Y_{i,j} = 1$ if x_i is labeled to have the j -th label, $Y_{i,j} = -1$ if x_i is labeled to not have the j -th label, and $Y_{i,j} = 0$ otherwise.

Let $\Delta(Y, H)$ be our loss function measuring the difference between Y and H . The multi-class volume approximation motivates the following family of transductive methods:

$$\min_{H \in \mathcal{H}_{P,Q}} \Delta(Y, H) + \gamma \cdot \frac{\text{tr}(H^\top Q H P)}{\|H\|_{\text{Fro}}^2},$$

where $\gamma > 0$ is a regularization parameter. The denominator $\|H\|_{\text{Fro}}^2$ is quite difficult to tackle, so we would like to eliminate it as El-Yaniv et al. (2008) and Niu et al. (2013a).

We fix a scale parameter $\tau > 0$, constrain H to be of norm τ , replace the feasible region $\mathcal{H}_{P,Q}$ with $\mathbb{R}^{n \times c}$ by extending the domain of $V(H)$ implicitly, and it becomes

$$\begin{aligned} \min_{H \in \mathbb{R}^{n \times c}} \quad & \Delta(Y, H) + \gamma \operatorname{tr}(H^\top QHP) \\ \text{s.t.} \quad & \|H\|_{\text{Fro}} = \tau. \end{aligned} \quad (8)$$

Although the optimization in (8) is done in $\mathbb{R}^{n \times c}$, the regularization is carried out relative to $\mathcal{H}_{P,Q}$, since under the constraint $\|H\|_{\text{Fro}} = \tau$, the regularization $\operatorname{tr}(H^\top QHP)$ is a weighted sum of the squares of cosines between $\operatorname{vec}(H)$ and the principal axes of $\mathcal{E}_{P,Q}$ like El-Yaniv et al. (2008).

Subsequently, we denote by $\mathbf{y}_1, \dots, \mathbf{y}_n$ and $\mathbf{r}_1, \dots, \mathbf{r}_n$ the c -dimensional vectors that satisfy $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top$ and $H = (\mathbf{r}_1, \dots, \mathbf{r}_n)^\top$. Consider the following loss functions to be $\Delta(Y, H)$ in optimization (8):

1. Squared losses over all data $\sum_{X_n} \|\mathbf{y}_i - \mathbf{r}_i\|_2^2$;
2. Squared losses over labeled data $\sum_{X_l} \|\mathbf{y}_i - \mathbf{r}_i\|_2^2$;
3. Linear losses over all data $\sum_{X_n} -\mathbf{y}_i^\top \mathbf{r}_i$;
4. Linear losses over labeled data $\sum_{X_l} -\mathbf{y}_i^\top \mathbf{r}_i$;

The first loss function has been used for multi-class transductive learning (Zhou et al., 2003) and the binary counterparts of the fourth and third loss functions have been used for binary transductive learning (El-Yaniv et al., 2008) and clustering (Niu et al., 2013a). Actually, the third and fourth loss functions are identical since \mathbf{y}_i for $x_i \in X_u$ is identically zero, and the first loss function is equivalent to them in (8) since $\sum_{X_n} \|\mathbf{y}_i\|_2^2$ and $\sum_{X_l} \|\mathbf{y}_i\|_2^2$ are constants and $\sum_{X_n} \|\mathbf{r}_i\|_2^2 = \tau^2$ is also a constant. The second loss function is undesirable for (8) due to an issue of the time complexity which will be discussed later. Thus, we instantiate $\Delta(Y, H) := \sum_{X_n} \|\mathbf{y}_i - \mathbf{r}_i\|_2^2 = \|Y - H\|_{\text{Fro}}^2$, and optimization (8) becomes

$$\begin{aligned} \min_{H \in \mathbb{R}^{n \times c}} \quad & \|Y - H\|_{\text{Fro}}^2 + \gamma \operatorname{tr}(H^\top QHP) \\ \text{s.t.} \quad & \|H\|_{\text{Fro}} = \tau. \end{aligned} \quad (9)$$

We refer to constrained optimization problem (9) as *multi-class approximate volume regularization* (MAVR). An unconstrained version of MAVR is then

$$\min_{H \in \mathbb{R}^{n \times c}} \|Y - H\|_{\text{Fro}}^2 + \gamma \operatorname{tr}(H^\top QHP). \quad (10)$$

4.2 Algorithm

Optimization (9) is non-convex, but we can rewrite it using the stacked soft response vector $\mathbf{h} = \operatorname{vec}(H)$ as

$$\begin{aligned} \min_{\mathbf{h} \in \mathbb{R}^{nc}} \quad & \|\mathbf{y} - \mathbf{h}\|_2^2 + \gamma \mathbf{h}^\top (P \otimes Q) \mathbf{h} \\ \text{s.t.} \quad & \|\mathbf{h}\|_2 = \tau, \end{aligned} \quad (11)$$

where $\mathbf{y} = \text{vec}(Y)$ is the vectorization of Y . In this representation, the objective is a second-degree polynomial and the constraint is an origin-centered sphere, and fortunately we could solve it exactly and efficiently following Forsythe and Golub (1965). To this end, a fundamental property of the Kronecker product is necessary (see, e.g., Theorems 13.10 and 13.12 of Laub, 2005):

Theorem 1. *Let $\lambda_{Q,1} \leq \dots \leq \lambda_{Q,n}$ be the eigenvalues and $\mathbf{v}_{Q,1}, \dots, \mathbf{v}_{Q,n}$ be the associated orthonormal eigenvectors of Q , $\lambda_{P,1} \leq \dots \leq \lambda_{P,c}$ and $\mathbf{v}_{P,1}, \dots, \mathbf{v}_{P,c}$ be those of P , and the eigen-decompositions of Q and P be $Q = V_Q \Lambda_Q V_Q^\top$ and $P = V_P \Lambda_P V_P^\top$. Then, the eigenvalues of $P \otimes Q$ are $\lambda_{P,j} \lambda_{Q,i}$ associated with orthonormal eigenvectors $\mathbf{v}_{P,j} \otimes \mathbf{v}_{Q,i}$ for $j = 1, \dots, c$, $i = 1, \dots, n$, and the eigen-decomposition of $P \otimes Q$ is $P \otimes Q = V_{PQ} \Lambda_{PQ} V_{PQ}^\top$, where $\Lambda_{PQ} = \Lambda_P \otimes \Lambda_Q$ and $V_{PQ} = V_P \otimes V_Q$.*

After we ignore the constants $\|\mathbf{y}\|_2^2$ and $\|\mathbf{h}\|_2^2$ in the objective of optimization (11), the Lagrange function is

$$\Phi(\mathbf{h}, \rho) = -2\mathbf{h}^\top \mathbf{y} + \gamma \mathbf{h}^\top (P \otimes Q) \mathbf{h} - \rho(\mathbf{h}^\top \mathbf{h} - \tau^2),$$

where $\rho \in \mathbb{R}$ is the Lagrangian multiplier for $\|\mathbf{h}\|_2^2 = \tau^2$. The stationary conditions are

$$\partial \Phi / \partial \mathbf{h} = -\mathbf{y} + \gamma(P \otimes Q) \mathbf{h} - \rho \mathbf{h} = \mathbf{0}_{nc}, \quad (12)$$

$$\partial \Phi / \partial \rho = \mathbf{h}^\top \mathbf{h} - \tau^2 = 0. \quad (13)$$

Hence, for any locally optimal solution (\mathbf{h}, ρ) where ρ/γ is not an eigenvalue of $P \otimes Q$, we have

$$\mathbf{h} = (\gamma P \otimes Q - \rho I_{nc})^{-1} \mathbf{y} \quad (14)$$

$$\begin{aligned} &= V_{PQ}(\gamma \Lambda_{PQ} - \rho I_{nc})^{-1} V_{PQ}^\top \mathbf{y} \\ &= (V_P \otimes V_Q)(\gamma \Lambda_{PQ} - \rho I_{nc})^{-1} \text{vec}(V_Q^\top Y V_P) \end{aligned} \quad (15)$$

based on Eq. (12) and Theorem 1. Next, we search for the feasible ρ for (12) and (13) which will lead to the globally optimal \mathbf{h} . Let $\mathbf{z} = \text{vec}(V_Q^\top Y V_P)$, then plugging (15) into (13) gives us

$$\mathbf{z}^\top (\gamma \Lambda_{PQ} - \rho I_{nc})^{-2} \mathbf{z} - \tau^2 = 0. \quad (16)$$

Let us sort the eigenvalues $\lambda_{P,1} \lambda_{Q,1}, \dots, \lambda_{P,c} \lambda_{Q,n}$ into a non-descending sequence $\{\lambda_{PQ,1}, \dots, \lambda_{PQ,nc}\}$, rearrange $\{z_1, \dots, z_{nc}\}$ accordingly, and find the smallest k_0 which satisfies $z_{k_0} \neq 0$. As a result, Eq. (16) implies that

$$g(\rho) = \sum_{k=k_0}^{nc} \frac{z_k^2}{(\gamma \lambda_{PQ,k} - \rho)^2} - \tau^2 = 0 \quad (17)$$

for any stationary ρ . By Theorem 4.1 of Forsythe and Golub (1965), the smallest root of $g(\rho)$ determines a unique \mathbf{h} so that (\mathbf{h}, ρ) is the globally optimal solution to $\Phi(\mathbf{h}, \rho)$, i.e., \mathbf{h} minimizes the objective of (11) globally. For this ρ , the only exception when it cannot determine \mathbf{h} by Eq. (14) is that ρ/γ is an eigenvalue of $P \otimes Q$, but this happens with probability zero. Finally, the theorem below points out the location of this ρ (the proof is in the appendix):

Algorithm 1 MAVR

Input: P, Q, Y, γ and τ

Output: H and ρ

- 1: Eigen-decompose P and Q ;
 - 2: Construct the function $g(\rho)$;
 - 3: Find the smallest root of $g(\rho)$;
 - 4: Recover \mathbf{h} using ρ and reshape \mathbf{h} to H .
-

Theorem 2. *The function $g(\rho)$ defined in Eq. (17) has exactly one root in the interval $[\rho_0, \gamma\lambda_{PQ,k_0})$ and no root in the interval $(-\infty, \rho_0)$, where $\rho_0 = \gamma\lambda_{PQ,k_0} - \|\mathbf{y}\|_2/\tau$.*

The algorithm of MAVR is summarized in Algorithm 1. It is easy to see that fixing $\rho = -1$ in Algorithm 1 instead of finding the smallest root of $g(\rho)$ suffices to solve optimization (10). Moreover, for a special case $P = I_c$ where I_c is the identity matrix of size c , any stationary H is simply

$$H = (\gamma Q - \rho I_n)^{-1} Y = V_Q(\gamma \Lambda_Q - \rho I_n)^{-1} V_Q^\top Y.$$

Let $\mathbf{z} = V_Q^\top Y \mathbf{1}_c$ where $\mathbf{1}_c$ is the all-one vector in \mathbb{R}^c , and k_0 is the smallest number that satisfies $z_{k_0} \neq 0$. Then the smallest root of

$$g(\rho) = \sum_{k=k_0}^n z_k^2 / (\gamma\lambda_{Q,k} - \rho)^2 - \tau^2$$

gives us the feasible ρ leading to the globally optimal H .

The asymptotic time complexity of Algorithm 1 is $O(n^3)$. More specifically, eigen-decomposing Q in the first step of Algorithm 1 costs $O(n^3)$, and this is the dominating computation time. Eigen-decomposing P just needs $O(c^3)$ and is negligible under the assumption that $n \gg c$ without loss of generality. In the second step, it requires $O(nc \log(nc))$ for sorting the eigenvalues of $P \otimes Q$ and $O(n^2c)$ for computing \mathbf{z} . Finding the smallest root of $g(\rho)$ based on a binary search algorithm uses $O(\log(\|\mathbf{y}\|_2))$ in the third step, and $\|\mathbf{y}\|_2 \leq \sqrt{l}$ for multi-class problems and $\|\mathbf{y}\|_2 \leq \sqrt{lc}$ for multi-label problems. In the final step, recovering \mathbf{h} is essentially same as computing \mathbf{z} and costs $O(n^2c)$.

We would like to comment a bit more on the asymptotic time complexity of MAVR. Firstly, we employ the squared losses over all data rather than the squared losses over labeled data. If the latter loss function was plugged in optimization (8), Eq. (14) would become

$$\mathbf{h} = (\gamma P \otimes Q - \rho I_{nc} + I_c \otimes J)^{-1} \mathbf{y},$$

where J is an n -by- n diagonal matrix such that $J_{i,i} = 1$ if x_i is labeled and $J_{i,i} = 0$ if x_i is unlabeled. The inverse in the expression above cannot be computed using the eigen-decompositions of P and Q , and hence the computational complexity would increase from $O(n^3)$ to $O(n^3c^3)$. Secondly, given fixed P and Q but different Y , γ , and τ , the computational complexity is $O(n^2c)$ if we reuse the eigen-decompositions of P and Q and the sorted eigenvalues of $P \otimes Q$. This property is especially advantageous for validating

and selecting hyperparameters. It is also quite useful for picking different $X_l \subset X_n$ to be labeled following transductive problem settings. Finally, the asymptotic time complexity $O(n^3)$ can hardly be improved based on existing techniques for optimizations (9) and (10). Even if ρ is fixed in optimization (10), the stationary condition Eq. (12) is a *discrete Sylvester equation* which consumes $O(n^3)$ for solving it (Sima, 1996).

4.3 Theoretical analyses

We provide two theoretical results. Under certain assumptions, the stability analysis upper bounds the difference of two optimal H and H' trained with two different label indicator matrices Y and Y' , and the error analysis bounds the difference of H from the ground truth.

Theorem 2 guarantees that $\rho < \gamma\lambda_{PQ,k_0}$. In fact, with high probability over the choice of Y , it holds that $k_0 = 1$ (we did not meet $k_0 > 1$ in our experiments). For this reason, we make the following assumption:

Fix P and Q , and allow Y to change according to the partition of X_n into different X_l and X_u . There is $C_{\gamma,\tau} > 0$, which just depends on γ and τ , such that for all optimal ρ trained with different Y , $\rho \leq \gamma\lambda_{PQ,1} - C_{\gamma,\tau}$.

Note that for unconstrained MAVR, there must be $C_{\gamma,\tau} > 1$ since $\gamma\lambda_{PQ,1} > 0$ and $\rho = -1$. Based on the above assumption and the lower bound of ρ in Theorem 2, we can prove the theorem below.

Theorem 3 (Stability of MAVR). *Assume the existence of $C_{\gamma,\tau}$. Let (H, ρ) and (H', ρ') be two globally optimal solutions trained with two different label indicator matrices Y and Y' respectively. Then,*

$$\|H - H'\|_{\text{Fro}} \leq \|Y - Y'\|_{\text{Fro}}/C_{\gamma,\tau} + |\rho - \rho'| \min\{\|Y\|_{\text{Fro}}, \|Y'\|_{\text{Fro}}\}/C_{\gamma,\tau}^2. \quad (18)$$

Consequently, for MAVR in optimization (9) we have

$$\|H - H'\|_{\text{Fro}} \leq \|Y - Y'\|_{\text{Fro}}/C_{\gamma,\tau} + \|Y\|_{\text{Fro}}\|Y'\|_{\text{Fro}}/\tau C_{\gamma,\tau}^2,$$

and for unconstrained MAVR in optimization (10) we have

$$\|H - H'\|_{\text{Fro}} \leq \|Y - Y'\|_{\text{Fro}}/C_{\gamma,\tau}.$$

In order to present an error analysis, we assume there is a ground-truth soft response matrix H^* with two properties. Firstly, the value of $V(H^*)$ should be bounded, namely,

$$V(H^*) = \frac{\text{tr}(H^{*\top}QH^*P)}{\|H^*\|_{\text{Fro}}^2} \leq C_h,$$

where $C_h > 0$ is a small number. This ensures that H^* lies in a large-volume region. Otherwise MAVR implementing the large volume principle can by no means learn some H close to H^* . Secondly, Y should contain certain information about H^* . MAVR makes

use of P , Q and Y only and the meanings of P and Q are fixed already, so MAVR may access the information about H^* only through Y . To make Y and H^* correlated, we assume that $Y = H^* + E$ where $E \in \mathbb{R}^{n \times c}$ is a noise matrix of the same size as Y and H^* . All entries of E are independent with zero mean, and the variance of them is σ_l or σ_u depending on its correspondence to a labeled or an unlabeled position in Y . We could expect that $\sigma_l \ll \sigma_u$, such that the entries of Y in labeled positions are close to the corresponding entries of H^* , but the entries of Y in unlabeled positions are completely corrupted and uninformative for recovering H^* . Notice that we need this generating mechanism of Y even if C_h/γ is the smallest eigenvalue of $P \otimes Q$, since $P \otimes Q$ may have multiple smallest eigenvalues and $\pm H$ have totally different meanings. Based on these assumptions, we can prove the theorem below.

Theorem 4 (Accuracy of MAVR). *Assume the existence of $C_{\gamma,\tau}$, C_h , and the generating process of Y from H^* and E . Let \tilde{l} and \tilde{u} be the numbers of the labeled and unlabeled positions in Y and assume that $\mathbb{E}_E \|Y\|_{\text{Fro}}^2 \leq \tilde{l}$ where the expectation is with respect to the noise matrix E . For each possible Y , let H be the globally optimal solution trained with it. Then,*

$$\begin{aligned} \mathbb{E}_E \|H - H^*\|_{\text{Fro}} &\leq (\sqrt{C_h} \gamma \lambda_{PQ,1} / C_{\gamma,\tau}) \|H^*\|_{\text{Fro}} \\ &\quad + \left(\max \left\{ \sqrt{\tilde{l}}/\tau - \gamma \lambda_{PQ,1} - 1, \gamma \lambda_{PQ,1} - C_{\gamma,\tau} + 1 \right\} / C_{\gamma,\tau} \right) \|H^*\|_{\text{Fro}} \\ &\quad + \sqrt{\tilde{l}\sigma_l^2 + \tilde{u}\sigma_u^2} / C_{\gamma,\tau} \end{aligned} \quad (19)$$

for MAVR in optimization (9), and

$$\mathbb{E}_E \|H - H^*\|_{\text{Fro}}^2 \leq (C_h/4) \|H^*\|_{\text{Fro}}^2 + \tilde{l}\sigma_l^2 + \tilde{u}\sigma_u^2 \quad (20)$$

for unconstrained MAVR in optimization (10).

The proofs of Theorems 3 and 4 are in the appendix. Considering the instability bounds in Theorem 3 and the error bounds in Theorem 4, unconstrained MAVR is superior to constrained MAVR in both cases. That being said, bounds are just bounds. We will demonstrate the potential of constrained MAVR in the next section by experiments.

5 Experiments

In this section, we numerically evaluate MAVR.

5.1 Serendipitous learning

We show how to handle serendipitous problems by MAVR directly without performing clustering (Hartigan and Wong, 1979; Ng et al., 2001; Sugiyama et al., 2014) or estimating the class-prior change (du Plessis and Sugiyama, 2012). The experimental results are

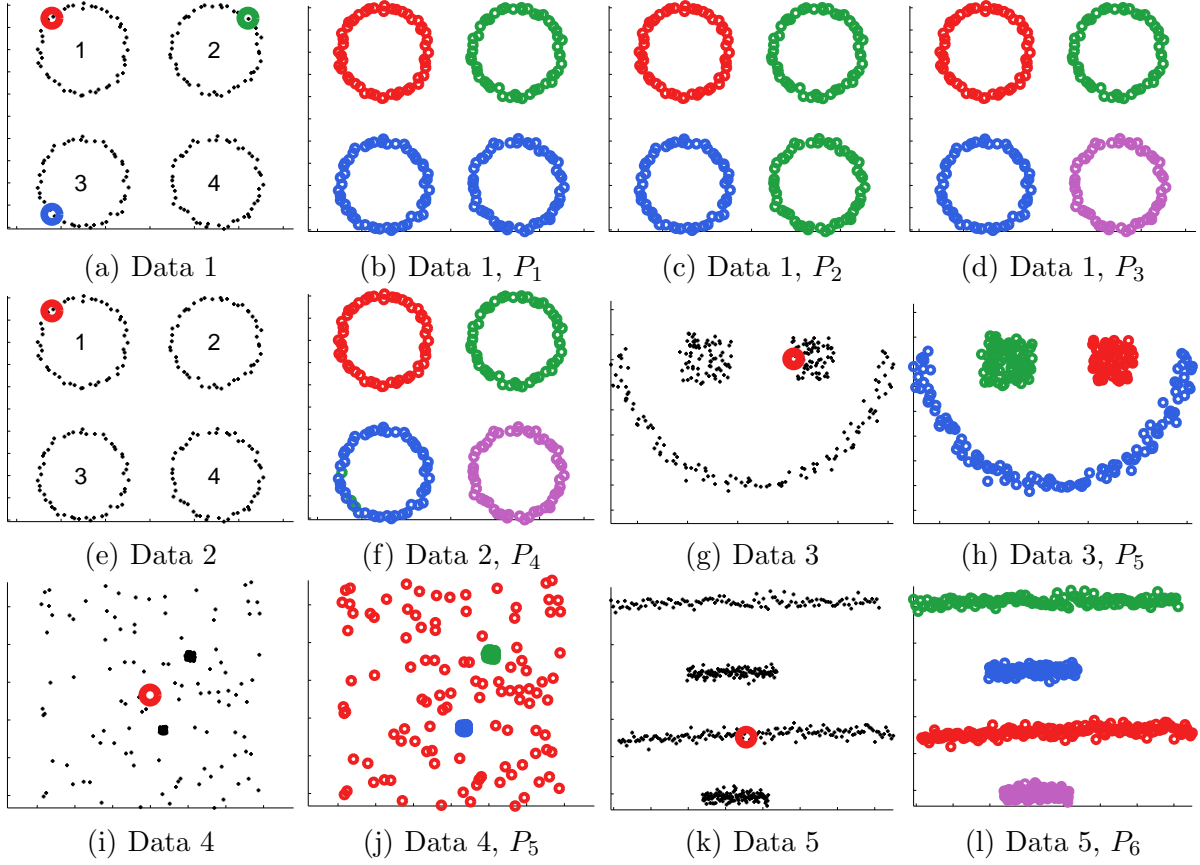


Figure 2: Experimental results of serendipitous learning problems on artificial data sets.

displayed in Figure 2. There are 5 data sets, and the latter 3 data sets are from Zelnik-Manor and Perona (2004). The matrix Q was specified as the *normalized graph Laplacian* (see, e.g., von Luxburg, 2007)¹ $L_{\text{nor}} = I_n - D^{-1/2}WD^{-1/2}$, where $W \in \mathbb{R}^{n \times n}$ is a similarity matrix and $D \in \mathbb{R}^{n \times n}$ is the degree matrix of W . The matrix P was specified by

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \end{pmatrix},$$

$$P_4 = \begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 2 & 0 & 1/2 \\ 1/2 & 0 & 2 & 1/2 \\ 1/2 & 1/2 & 1/2 & 3 \end{pmatrix}, P_5 = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}, P_6 = \begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{pmatrix}.$$

For data sets 1 and 2 we used the Gaussian similarity

$$W_{i,j} = \exp(-\|x_i - x_j\|_2^2 / (2\sigma^2))$$

¹Though the graph Laplacian matrices have zero eigenvalues, they would not cause algorithmic problems when used as Q .

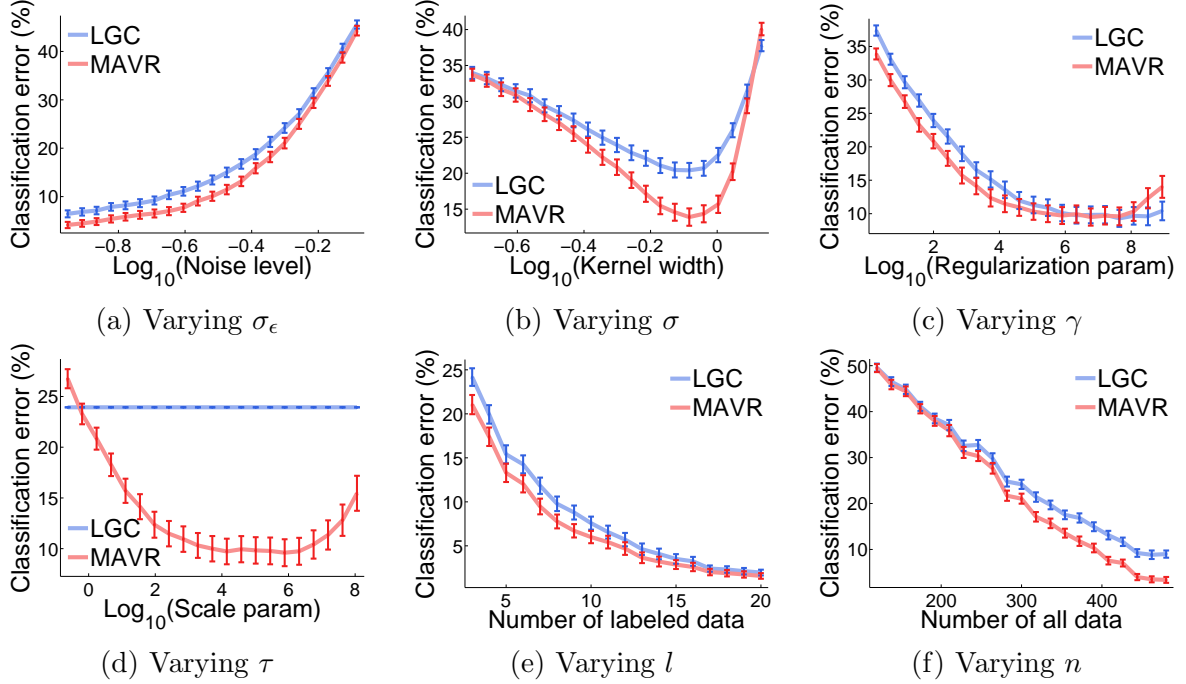


Figure 3: Experimental results on the artificial data set 3circles. Means with standard errors are shown.

with the kernel width $\sigma = 0.25$, and for data sets 3 to 5 we applied the local-scaling similarity (Zelnik-Manor and Perona, 2004)

$$W_{i,j} = \exp(-\|x_i - x_j\|_2^2 / (2\sigma_i\sigma_j)), \quad \sigma_i = \|x_i - x_i^{(k)}\|_2$$

with the number of nearest neighbors $k = 7$, where $x_i^{(k)}$ is the k -th nearest neighbor of x_i in X_n . We set $\gamma = 99$ and $\tau = \sqrt{l}$. Furthermore, a class balance regularization was imposed for data sets 2 to 5. The detail is omitted here due to the space limit, while the idea is to encourage balanced total responses of all c classes. For this regularization, the regularization parameter was $\gamma' = 1$. We can see that in Figure 2, MAVR successfully classified the data belonging to the known classes and simultaneously clustered the data belonging to the unknown classes. By specifying different P , we could control the influence of the known classes on the unknown classes.

5.2 Multi-class learning

A state-of-the-art multi-class transductive learning method named *learning with local and global consistency* (LGC) (Zhou et al., 2003) is closely related to MAVR. Actually, if we specify $P = I_c$ and $Q = L_{\text{nor}}$, unconstrained MAVR will be reduced to LGC exactly. Although LGC is motivated by the label propagation viewpoint, it can be written as optimization (4) in Zhou et al. (2003). Here, we illustrate the nuance of constrained MAVR and LGC that is unconstrained MAVR using an artificial data set.

The artificial data set *3circles* is generated as follows. We have three classes with the class ratio $1/6, 1/3, 1/2$. Let y_i be the ground-truth label of x_i , then x_i is generated by

$$x_i = (6y_i \cos(a_i) + \epsilon_{i,1}, 5y_i \sin(a_i) + \epsilon_{i,2})^\top \in \mathbb{R}^2,$$

where a_i is an angle drawn i.i.d. from the uniform distribution $\mathcal{U}(0, 2\pi)$, and $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are noises drawn i.i.d. from the normal distribution $\mathcal{N}(0, \sigma_\epsilon^2)$. We vary one factor and fix all other factors. The default values of these factors are $\sigma_\epsilon = 0.5$, $\sigma = 0.5$, $l = 3$, $n = 300$, $\gamma = 99$, and $\tau = \sqrt{l}$. Figure 3 shows the experimental results, where the means with the standard errors of the classification error rates are plotted. For each task that corresponds to a full specification of all factors, MAVR and LGC were repeatedly ran on 100 random samplings. We can see from Figure 3 that the performance of LGC was usually not as good as MAVR.

Over the past decades, a huge number of transductive learning and semi-supervised learning methods have been proposed based on various motivations as graph cut (Blum and Chawla, 2001), random walk (Zhu et al., 2003), manifold regularization (Belkin et al., 2006), and information maximization (Niu et al., 2013b), just to name a few. A state-of-the-art semi-supervised learning method called *Laplacian regularized least squares* (LapRLS) (Belkin et al., 2006) is included to be compared with MAVR besides LGC.

The experimental results are reported in Figure 4. Similarly to Figure 3, the means with the standard errors of the classification error rates are shown where 4 methods were repeatedly ran on 100 random samplings for each task. We considered another specification of Q as the *unnormalized graph Laplacian* $L_{\text{un}} = D - W$ which was also employed by LapRLS. The cosine similarity is defined by

$$W_{i,j} = x_i^\top x_j / \|x_i\|_2 \|x_j\|_2 \text{ if } x_i \sim_k x_j, \quad W_{i,j} = 0 \text{ otherwise,}$$

where $x_i \sim_k x_j$ means x_i and x_j are among the k -nearest neighbors of each other. We set $l = n/10$ for all involved n in Figure 4, and there seems no reliable model selection method given very few labeled data, so we select the best hyperparameters for each method using the labels of unlabeled data from 10 additional random samplings. Specifically, σ is the median distance $\times \{1/16, 1/8, 1/4, 1/2, 1\}$, and k is from $\{1, 3, 5, 7, 9\}$ for both local-scaling and cosine similarities; τ is $\sqrt{l} \times \{1/16, 1/8, 1/4, 1/2, 1\}$. The hyperparameters are all fixed since it resulted in more stable performance. For MAVR, LGC, and λ_I of LapRLS, it was fixed to 99 if the Gaussian and cosine similarities were used and 1 if the local-scaling similarity was used; λ_A of LapRLS was 10^{-3} if the Gaussian and local-scaling similarities were used and 10^3 if the cosine similarity was used since LapRLS also needed W that was too sparse and near singular, but an exception was panel (i) where $\lambda_A = 10^{-3}$ gave lower error rates of LapRLS. We can see from Figure 4 that two MAVR methods often compared favorably with the state-of-the-art methods LGC and LapRLS, which implies that our proposed multi-class volume approximation is reasonable and practical.

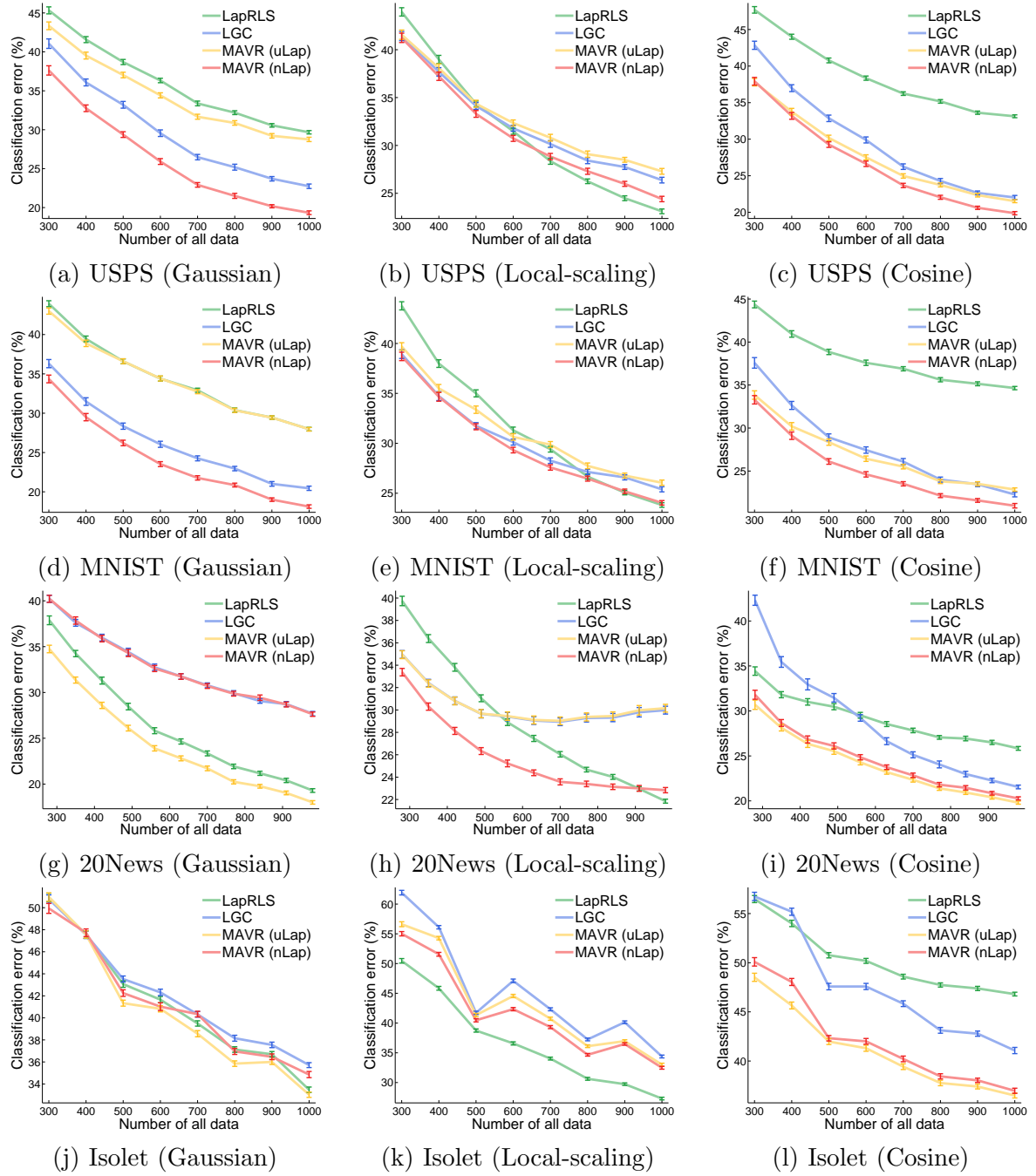


Figure 4: Experimental results on USPS, MNIST, 20Newsgroups and Isolet. Means with standard errors are shown.

6 Conclusions

We proposed a multi-class volume approximation that can be applied to several transductive problem settings such as multi-class, multi-label and serendipitous learning. The

resultant learning method is non-convex in nature but can be solved exactly and efficiently. It is theoretically justified by our stability and error analyses and experimentally demonstrated promising.

A Proofs

A.1 Proof of Theorem 2

The derivative of $g(\rho)$ is

$$g'(\rho) = \sum_{k=k_0}^{nc} \frac{2z_k^2}{(\gamma\lambda_{PQ,k} - \rho)^3} - \tau^2.$$

Hence, $g'(\rho) > 0$ whenever $\rho < \gamma\lambda_{PQ,k_0}$, and $g(\rho)$ is strictly increasing in the interval $(-\infty, \gamma\lambda_{PQ,k_0})$. Moreover,

$$\lim_{\rho \rightarrow -\infty} g(\rho) = -\tau^2 \quad \text{and} \quad \lim_{\rho \rightarrow \gamma\lambda_{PQ,k_0}} g(\rho) = +\infty,$$

and thus $g(\rho)$ has exactly one root in $(-\infty, \gamma\lambda_{PQ,k_0})$. Notice that $\|\mathbf{z}\|_2 = \|\text{vec}(V_Q^\top Y V_P)\|_2 = \|V_{PQ}^\top \mathbf{y}\|_2 = \|\mathbf{y}\|_2$ since V_{PQ} is an orthonormal matrix, and then $\rho_0 = \gamma\lambda_{PQ,k_0} - \|\mathbf{y}\|_2/\tau = \gamma\lambda_{PQ,k_0} - \|\mathbf{z}\|_2/\tau$. As a result,

$$\begin{aligned} g(\rho_0) &= \sum_{k=k_0}^{nc} \frac{z_k^2}{(\gamma\lambda_{PQ,k} - \rho_0)^2} - \tau^2 \\ &= \sum_{k=k_0}^{nc} \frac{z_k^2}{(\gamma\lambda_{PQ,k} - \gamma\lambda_{PQ,k_0} + \|\mathbf{z}\|_2/\tau)^2} - \tau^2 \\ &\leq \sum_{k=k_0}^{nc} \frac{z_k^2}{(\|\mathbf{z}\|_2/\tau)^2} - \tau^2 \\ &= \left(\frac{\sum_{k=k_0}^{nc} z_k^2}{\|\mathbf{z}\|_2^2} - 1 \right) \tau^2 \\ &\leq 0, \end{aligned}$$

where the first inequality is because $\lambda_{PQ,k} \geq \lambda_{PQ,k_0}$ for $k \geq k_0$. The fact that $g(\rho_0) \leq 0$ concludes that the only root in $(-\infty, \gamma\lambda_{PQ,k_0})$ is in $[\rho_0, \gamma\lambda_{PQ,k_0})$ but not $(-\infty, \rho_0)$. \square

A.2 Proof of Theorem 3

Denote by $\mathbf{h} = \text{vec}(H)$, $\mathbf{y} = \text{vec}(Y)$ and $M = (\gamma P \otimes Q - \rho I_{nc})$, and denote by \mathbf{h}' , \mathbf{y}' and M' similarly. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be two functions extracting the smallest and largest eigenvalues of a matrix. Under our assumption,

$$\lambda_{\min}(M) = \gamma\lambda_{PQ,1} - \rho \geq C_{\gamma,\tau} > 0$$

which means that M is positive definite, and so is M' . By Eq. (14),

$$\begin{aligned}\mathbf{h} - \mathbf{h}' &= M^{-1}\mathbf{y} - M'^{-1}\mathbf{y}' \\ &= M^{-1}(\mathbf{y} - \mathbf{y}') + (M^{-1} - M'^{-1})\mathbf{y}' \\ &= M^{-1}(\mathbf{y} - \mathbf{y}') + M^{-1}(M' - M)M'^{-1}\mathbf{y}' \\ &= M^{-1}(\mathbf{y} - \mathbf{y}') + (\rho' - \rho)M^{-1}M'^{-1}\mathbf{y}'.\end{aligned}$$

Note that $\|A\mathbf{v}\|_2 \leq \lambda_{\max}(A)\|\mathbf{v}\|_2$ for any symmetric positive-definite matrix A and any vector \mathbf{v} , as well as $\lambda_{\max}(AB) \leq \lambda_{\max}(A)\lambda_{\max}(B)$ for any symmetric positive-definite matrices A and B . Hence,

$$\begin{aligned}\|\mathbf{h} - \mathbf{h}'\|_2 &= \|M^{-1}(\mathbf{y} - \mathbf{y}') + (\rho' - \rho)M^{-1}M'^{-1}\mathbf{y}'\|_2 \\ &\leq \|M^{-1}(\mathbf{y} - \mathbf{y}')\|_2 + |\rho - \rho'| \|M^{-1}M'^{-1}\mathbf{y}'\|_2 \\ &\leq \lambda_{\max}(M^{-1})\|\mathbf{y} - \mathbf{y}'\|_2 + \lambda_{\max}(M^{-1})\lambda_{\max}(M'^{-1})|\rho - \rho'| \|\mathbf{y}'\|_2 \\ &\leq \frac{\|\mathbf{y} - \mathbf{y}'\|_2}{C_{\gamma,\tau}} + \frac{|\rho - \rho'| \|\mathbf{y}'\|_2}{C_{\gamma,\tau}^2},\end{aligned}$$

where the first inequality is the triangle inequality, the second inequality is because M^{-1} and M'^{-1} are symmetric positive definite, and the third inequality follows from $\lambda_{\max}(M^{-1}) = 1/\lambda_{\min}(M)$ and $\lambda_{\max}(M'^{-1}) = 1/\lambda_{\min}(M')$. Due to the symmetry of \mathbf{h} and \mathbf{h}' ,

$$\|\mathbf{h} - \mathbf{h}'\|_2 \leq \frac{\|\mathbf{y} - \mathbf{y}'\|_2}{C_{\gamma,\tau}} + \frac{|\rho - \rho'| \min\{\|\mathbf{y}\|_2, \|\mathbf{y}'\|_2\}}{C_{\gamma,\tau}^2}.$$

This inequality is the vectorization of (18).

For MAVR in optimization (9), Theorem 2 together with our assumption indicates that

$$\begin{aligned}\gamma\lambda_{PQ,1} - \|\mathbf{y}\|_2/\tau &\leq \rho < \gamma\lambda_{PQ,1}, \\ \gamma\lambda_{PQ,1} - \|\mathbf{y}'\|_2/\tau &\leq \rho' < \gamma\lambda_{PQ,1},\end{aligned}$$

so $|\rho' - \rho| \leq \max\{\|\mathbf{y}\|_2/\tau, \|\mathbf{y}'\|_2/\tau\}$ and

$$\begin{aligned}\|\mathbf{h} - \mathbf{h}'\|_2 &\leq \frac{\|\mathbf{y} - \mathbf{y}'\|_2}{C_{\gamma,\tau}} + \frac{\max\{\|\mathbf{y}\|_2, \|\mathbf{y}'\|_2\} \min\{\|\mathbf{y}\|_2, \|\mathbf{y}'\|_2\}}{\tau C_{\gamma,\tau}^2} \\ &= \frac{\|\mathbf{y} - \mathbf{y}'\|_2}{C_{\gamma,\tau}} + \frac{\|\mathbf{y}\|_2 \|\mathbf{y}'\|_2}{\tau C_{\gamma,\tau}^2}.\end{aligned}$$

For unconstrained MAVR in optimization (10), we have

$$\|\mathbf{h} - \mathbf{h}'\|_2 \leq \frac{\|\mathbf{y} - \mathbf{y}'\|_2}{C_{\gamma,\tau}},$$

since $\rho = \rho' = -1$. □

A.3 Proof of Theorem 4

Denote by $\mathbf{h} = \text{vec}(H)$, $\mathbf{y} = \text{vec}(Y)$, $\mathbf{h}^* = \text{vec}(H^*)$, $\mathbf{e} = \text{vec}(E)$, and $M = \gamma P \otimes Q$. The Kronecker product $P \otimes Q$ is symmetric and positive definite, and then $M^{1/2}$ is a well-defined symmetric and positive-definite matrix. We can know based on $V(H^*) \leq C_h$ that

$$\|M^{1/2}\mathbf{h}^*\|_2 = \sqrt{\gamma\mathbf{h}^{*\top}(P \otimes Q)\mathbf{h}^*} \leq \sqrt{\gamma C_h \|\mathbf{h}^*\|_2^2} = \sqrt{\gamma C_h} \|\mathbf{h}^*\|_2.$$

Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be two functions extracting the smallest and largest eigenvalues of a matrix. In the following, we will frequently use that $\|A\mathbf{v}\|_2 \leq \lambda_{\max}(A)\|\mathbf{v}\|_2$ for any symmetric positive-definite matrix A and any vector \mathbf{v} .

Consider unconstrained MAVR in optimization (10) first. Since $\rho = -1$,

$$\begin{aligned} \mathbf{h} - \mathbf{h}^* &= (M + I_{nc})^{-1}\mathbf{y} - \mathbf{h}^* \\ &= (M + I_{nc})^{-1}(\mathbf{h}^* + \mathbf{e}) - (M + I_{nc})^{-1}(M + I_{nc})\mathbf{h}^* \\ &= -(M + I_{nc})^{-1}M\mathbf{h}^* + (M + I_{nc})^{-1}\mathbf{e}. \end{aligned}$$

As a consequence,

$$\mathbb{E}\|\mathbf{h} - \mathbf{h}^*\|_2^2 = \|(M + I_{nc})^{-1}M\mathbf{h}^*\|_2^2 + \mathbb{E}\|(M + I_{nc})^{-1}\mathbf{e}\|_2^2,$$

since $\mathbb{E}[(M + I_{nc})^{-1}\mathbf{e}] = (M + I_{nc})^{-1}\mathbb{E}\mathbf{e} = \mathbf{0}_{nc}$. Subsequently,

$$\begin{aligned} \|(M + I_{nc})^{-1}M\mathbf{h}^*\|_2 &\leq \lambda_{\max}((M + I_{nc})^{-1}M^{1/2}) \cdot \|M^{1/2}\mathbf{h}^*\|_2 \\ &\leq \lambda_{\max}((\gamma P \otimes Q + I_{nc})^{-1}(\gamma P \otimes Q)^{1/2}) \cdot \sqrt{\gamma C_h} \|\mathbf{h}^*\|_2 \\ &= \sqrt{\gamma C_h} \lambda_{\max}\left(\frac{\sqrt{\gamma}}{\gamma + 1}(\Lambda_{PQ} + I_{nc})^{-1}\Lambda_{PQ}^{1/2}\right) \|\mathbf{h}^*\|_2 \\ &\leq \sqrt{C_h} \lambda_{\max}((\Lambda_{PQ} + I_{nc})^{-1}\Lambda_{PQ}^{1/2}) \|\mathbf{h}^*\|_2 \\ &\leq \frac{1}{2}\sqrt{C_h} \|\mathbf{h}^*\|_2, \end{aligned}$$

where the last inequality is because the eigenvalues of $(\Lambda_{PQ} + I_{nc})^{-1}\Lambda_{PQ}^{1/2}$ are $\frac{\sqrt{\lambda_{PQ,1}}}{\lambda_{PQ,1}+1}, \dots, \frac{\sqrt{\lambda_{PQ,nc}}}{\lambda_{PQ,nc}+1}$ and

$$\sup_{\lambda \geq 0} \frac{\sqrt{\lambda}}{\lambda + 1} = \frac{1}{2}.$$

On the other hand,

$$\begin{aligned} \mathbb{E}\|(M + I_{nc})^{-1}\mathbf{e}\|_2^2 &\leq (\lambda_{\max}((M + I_{nc})^{-1}))^2 \cdot \mathbb{E}\|\mathbf{e}\|_2^2 \\ &= \frac{\mathbb{E}[\mathbf{e}^\top \mathbf{e}]}{(\lambda_{\min}(M + I_{nc}))^2} \\ &\leq \tilde{l}\sigma_t^2 + \tilde{u}\sigma_u^2. \end{aligned}$$

Hence,

$$\mathbb{E}\|\mathbf{h} - \mathbf{h}^*\|_2^2 \leq \frac{1}{4}C_h\|\mathbf{h}^*\|_2^2 + \tilde{l}\sigma_l^2 + \tilde{u}\sigma_u^2,$$

which completes the proof of inequality (20).

Next, consider MAVR in optimization (9). We would have

$$\begin{aligned}\mathbf{h} - \mathbf{h}^* &= (M - \rho I_{nc})^{-1}\mathbf{y} - \mathbf{h}^* \\ &= (M - \rho I_{nc})^{-1}(\mathbf{h}^* + \mathbf{e}) - (M - \rho I_{nc})^{-1}(M - \rho I_{nc})\mathbf{h}^* \\ &= -(M - \rho I_{nc})^{-1}(M - (\rho + 1)I_{nc})\mathbf{h}^* + (M - \rho I_{nc})^{-1}\mathbf{e}.\end{aligned}$$

In general, $\mathbb{E}[(M - \rho I_{nc})^{-1}\mathbf{e}] \neq \mathbf{0}_{nc}$ since ρ depends on \mathbf{e} . Furthermore, $M - (\rho + 1)I_{nc}$ may have negative eigenvalues when $\gamma\lambda_{PQ,1} - 1 < \rho \leq \gamma\lambda_{PQ,1} - C_{\gamma,\tau}$. Taking the expectation of $\|\mathbf{h} - \mathbf{h}^*\|_2$,

$$\begin{aligned}\mathbb{E}\|\mathbf{h} - \mathbf{h}^*\|_2 &\leq \mathbb{E}\|(M - \rho I_{nc})^{-1}(M - (\rho + 1)I_{nc})\mathbf{h}^*\|_2 + \mathbb{E}\|(M - \rho I_{nc})^{-1}\mathbf{e}\|_2 \\ &\leq \mathbb{E}\|(M - \rho I_{nc})^{-1}M\mathbf{h}^*\|_2 + \mathbb{E}[|\rho + 1|]\|(M - \rho I_{nc})^{-1}\mathbf{h}^*\|_2 + \mathbb{E}\|(M - \rho I_{nc})^{-1}\mathbf{e}\|_2.\end{aligned}$$

Subsequently,

$$\begin{aligned}\mathbb{E}\|(M - \rho I_{nc})^{-1}M\mathbf{h}^*\|_2 &\leq \sup_{\rho} \lambda_{\max}((M - \rho I_{nc})^{-1}M^{1/2}) \cdot \sqrt{\gamma C_h}\|\mathbf{h}^*\|_2 \\ &= \sup_{\rho} \sqrt{C_h} \lambda_{\max}\left((\Lambda_{PQ} - \rho/\gamma I_{nc})^{-1}\Lambda_{PQ}^{1/2}\right)\|\mathbf{h}^*\|_2 \\ &\leq \sqrt{C_h}\|\mathbf{h}^*\|_2 \cdot \sup_{\rho \leq \gamma\lambda_{PQ,1} - C_{\gamma,\tau}} \sup_{\lambda \geq \lambda_{PQ,1}} \left(\frac{\sqrt{\lambda}}{\lambda - \rho/\gamma}\right) \\ &\leq \frac{\sqrt{C_h}\gamma\lambda_{PQ,1}}{C_{\gamma,\tau}}\|\mathbf{h}^*\|_2.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}[|\rho + 1|]\|(M - \rho I_{nc})^{-1}\mathbf{h}^*\|_2 &\leq \mathbb{E}|\rho + 1| \cdot \sup_{\rho} \lambda_{\max}((M - \rho I_{nc})^{-1})\|\mathbf{h}^*\|_2 \\ &\leq \frac{\|\mathbf{h}^*\|_2}{C_{\gamma,\tau}} \cdot \mathbb{E} \max\{-\rho - 1, \sup_{\rho} \rho + 1\} \\ &\leq \frac{\|\mathbf{h}^*\|_2}{C_{\gamma,\tau}} \cdot \max\{\mathbb{E}\|\mathbf{y}\|_2/\tau - \gamma\lambda_{PQ,1} - 1, \gamma\lambda_{PQ,1} - C_{\gamma,\tau} + 1\} \\ &= \frac{\|\mathbf{h}^*\|_2}{C_{\gamma,\tau}} \cdot \max\{\sqrt{\tilde{l}}/\tau - \gamma\lambda_{PQ,1} - 1, \gamma\lambda_{PQ,1} - C_{\gamma,\tau} + 1\}.\end{aligned}$$

where we used the fact that $\sup_{\rho} \rho$ is independent of \mathbf{e} , and applied *Jensen's inequality* to obtain that

$$\mathbb{E}\|\mathbf{y}\|_2 \leq \sqrt{\mathbb{E}\|\mathbf{y}\|_2^2} \leq \sqrt{\tilde{l}}.$$

In the end,

$$\begin{aligned}
\mathbb{E}\|(M - \rho I_{nc})^{-1} \mathbf{e}\|_2 &\leq \sup_{\rho} \lambda_{\max}((M - \rho I_{nc})^{-1}) \cdot \mathbb{E}\|\mathbf{e}\|_2 \\
&\leq \frac{\mathbb{E}\sqrt{\mathbf{e}^\top \mathbf{e}}}{C_{\gamma, \tau}} \\
&\leq \frac{\sqrt{\mathbb{E}[\mathbf{e}^\top \mathbf{e}]}}{C_{\gamma, \tau}} \\
&= \frac{\sqrt{\tilde{l}\sigma_l^2 + \tilde{u}\sigma_u^2}}{C_{\gamma, \tau}},
\end{aligned}$$

where the third inequality is due to Jensen's inequality. Therefore, inequality (19) follows by combining the three upper bounds of expectations. \square

References

- M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.
- A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, 2001.
- M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *ICML*, 2012.
- R. El-Yaniv, D. Pechyony, and V. Vapnik. Large margin vs. large volume in transductive learning. *Machine Learning*, 72(3):173–188, 2008.
- G. Forsythe and G. Golub. On the stationary values of a second-degree polynomial on the unit sphere. *Journal of the Society for Industrial and Applied Mathematics*, 13(4):1050–1068, 1965.
- J. A. Hartigan and M. A. Wong. A k -means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.
- X. Kong, M. Ng, and Z.-H. Zhou. Transductive multi-label learning via label set propagation. *IEEE Transaction on Knowledge and Data Engineering*, 25(3):704–719, 2013.
- A. J. Laub. *Matrix Analysis for Scientists and Engineers*. Society for Industrial and Applied Mathematics, 2005.
- S. Li and W. Ng. Maximum volume outlier detection and its applications in credit risk analysis. *International Journal on Artificial Intelligence Tools*, 22(5), 2013.

- A. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.
- G. Niu, B. Dai, L. Shang, and M. Sugiyama. Maximum volume clustering: A new discriminative clustering approach. *Journal of Machine Learning Research*, 14:2641–2687, 2013a.
- G. Niu, W. Jitkrittum, B. Dai, H. Hachiya, and M. Sugiyama. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *ICML*, 2013b.
- V. Sima. *Algorithms for Linear-Quadratic Optimization*. Marcel Dekker, 1996.
- M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 26(1):84–131, 2014.
- M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. In *NIPS*, 2001.
- V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 1982.
- V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
- U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.
- D. Zhang, Y. Liu, and L. Si. Serendipitous learning: Learning beyond the predefined label space. In *KDD*, 2011.
- D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.
- X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.